DOCUMENT RESUME

ED 233 638 HE 016 426

| | |
|---|---|
| AUTHOR | Perry, Raymond P. |
| TITLE | Educational Seduction: Implications for the Evaluation and Improvement of Teaching. Report #7. |
| INSTITUTION | British Columbia Univ., Vancouver. Centre for the Improvement of Teaching and Evaluation. |
| REPORT NO | CITE-7 |
| PUB DATE | Jun 81 |
| NOTE | 48p.; Paper presented at a colloquium sponsored by the Standing Committee on Teaching and the Center for the Improvement of Teaching and Evaluation, University of British Columbia (Vancouver, British Columbia, March 1981). |
| AVAILABLE FROM | Centre for the Improvement of Teaching and Evaluation, University of British Columbia, Faculty of Education, Vancouver, British Columbia ($2.00 Canadian). |
| PUB TYPE | Speeches/Conference Papers (150) -- Information Analyses (070) |
| EDRS PRICE | MF01/PC02 Plus Postage. |
| DESCRIPTORS | Educational Research; *Evaluation Criteria; *Faculty Evaluation; Higher Education; Instructional Improvement; *Research Design; *Student Evaluation of Teacher Performance; Teacher Behavior; *Teacher Effectiveness; *Teaching Styles |

ABSTRACT

Issues pertaining to the validity of student ratings of teacher effectiveness, research on "educational seduction," and implications of the research for the evaluation and improvement of teaching are reviewed. Criteria that have been used to validate student ratings include student achievement, faculty ratings, self-ratings, research productivity, and teacher's educational background. The concept of educational seduction, also called the "Dr. Fox Effect," comes from the claim that an entertaining, charismatic instructor can receive favorable student ratings, while offering insufficient lecture content. To test this hypothesis, an experimental design is needed involving two independent variables (instructor expressiveness and lecture content) and two dependent variables (student ratings and achievement). Different research approaches are described, including those employed by Williams and Ware and by Perry, Abrami, and Leventhal. William and Ware argue that their research supports educational seduction, and that the effect can be generalized to classrooms in the field setting. Perry, Abrami, and Leventhal claim that the educational seduction effect is not supported by their empirical results. (SW)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED233638

# EDUCATIONAL SEDUCTION: IMPLICATIONS FOR THE EVALUATION AND IMPROVEMENT OF TEACHING

Raymond P. Perry
Instructional Research Laboratory
Department of Psychology
University of Manitoba

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Raymond Perry

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

✓ This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

RE016426

Report # 7

© June 1981

Colloquium presented at the University of British Columbia, Vancouver, Canada, March 1981, sponsored by the Standing Committee on Teaching (SCOT) and Center for the Improvement of Teaching and Evaluation (CITE) of the Faculty of Education.

## PREFACE

The Standing Committee on Teaching was pleased to sponsor a visit to U.B.C. of Dr. Raymond P. Perry on March 4 and 5, 1981. Dr. Perry, from the University of Manitoba, consulted with staff, held a graduate seminar and a conversation hour with faculty and delivered a colloquium address which forms the body of this report.

Consistent with the purposes of the Standing Committee and the Centre for the Improvement of Teaching (CITE), Dr. Perry's paper is addressed to professors who are concerned with course improvement. In it he reports the results of an extensive study of the phenomenon known as the "Dr. Fox effect" or educational seduction. Implications for teaching evaluation and improvement and for further study based upon these findings are drawn.

We are pleased to present Dr. Perry's paper, a report on an important educational research project, in this format as the seventh CITE report. Dr. Perry delivered the fourth annual colloquium in the series begun exactly three years previously by Dr. Patricia A. Cranton of McGill University and continued in 1979 by Dr. Harry G. Murray of the University of Western Ontario and in 1980 by Dr. Janet G. Donald of McGill University. The previous three colloquia concerned with teaching evaluation are Reports #1, #5 and #6 in this series respectively.

The author and sponsors of this report invite comments, questions or remarks from recipient readers. Address these to Dr. Perry at the Department of Psychology, the University of Manitoba, Winnipeg, Manitoba, or to CITE at the U.B.C. Faculty of Education, Vancouver, B.C. Additional copies of this or other reports in the series are available from CITE upon request (enclose $2.00 Canadian for each copy desired to defray production and postage expenses).

Stephen F. Foster, J. Gordon Nelson

Vancouver, B.C.

March 1981

# EDUCATIONAL SEDUCTION: IMPLICATIONS FOR THE EVALUATION AND IMPROVEMENT OF TEACHING[1]

Raymond P. Perry[2]
University of Manitoba

## I. THE EVALUATION OF INSTRUCTION

### A. Student Ratings

During the past fifteen years increasing attention has been directed toward teaching effectiveness in post-secondary education. Attempts to improve the overall quality of instruction have focused on the development of appropriate evaluation procedures. Student ratings appear to be one of the most widely-adopted procedures for measuring teaching effectiveness. Gage (1958) suggests that student ratings are used for at least three purposes: (1) to improve the quality of instruction by providing feedback to the instructor; (2) to aid administrative decision-making with regards to tenure, promotion, and merit, and (3) to provide a criterion measure for research on teaching. Werdell (1967) added a fourth purpose to Gage's list: to identify good teachers for students wishing to enhance the quality of their education.

Proponents of student ratings argue that students are well-suited to evaluate teaching. Having participated in the educational system at least twelve years prior to attending college, and having been exposed to thousands of hours of instruction and dozens of teachers, it seems reasonable to assume that students develop some concept of

instruction, and can discriminate between teachers based on that concept. Accordingly, students should be reasonably accurate judges of teaching ability. It is worth noting that the kinds of discriminations required of the students are important. Discriminations between broad categories of teaching are easier to make, and are more likely to be accurate, than fine distinctions. For example, a five-point rating scale which requires the student to distinguish between poor, average, and good teaching may be more useful than a highly-refined scale which requires the student to discriminate between ten or more categories of teaching.

Critics contend, however, that students are not able to evaluate accurately the many facets of teaching. Instruction judged by students as "irrelevant digression" and "needless complexity" may be viewed by colleagues as "important elaborations" and "necessary specificity". The critics maintain that numerous extraneous factors bias students' assessments such as: the effects of instructor personality characteristics on student judgments; the use of grades by instructors to manipulate ratings; the stability and consistency of student opinions. These and many other arguments have been used to condemn the use of student ratings in post-secondary institutions. They have also contributed to an extensive research literature concerned with whether student ratings are reliable and valid.

In eliminating student ratings, many critics believe that the problem has been resolved because unreliable and invalid information will no longer be used. The logical fallacy here is the assumption that the absence of <u>formal</u> evaluation procedures implies no evaluation. Recent research evidence, however, suggests that there are active, <u>informal</u>, word-of-mouth evaluation networks operating in education institutions,

coincidently with, or in the absence of, formal evaluation procedures. Leventhal, Abrami, Perry, and Breen (1975) investigated students' reasons for section-choice in large, multisectioned psychology courses. The two most significant reasons for selecting a particular section were class time and instructor reputation, with the latter having significantly greater importance for senior students. Perry, Abrami, Leventhal, and Check (1979) found that an instructor's reputation based on previous student evaluations can affect subsequent student ratings and, to a lesser extent, student achievement. Students presented with either a negative or positive instructor reputation, rated the instructor in accordance with their initial expectations, even though all students saw the same lecture. It would appear therefore, that informal evaluation procedures may operate independently of formal procedures, and that they can influence various judgments about effective teaching.

Accordingly, it is necessary to consider the strengths and weaknesses of student ratings in comparison to other formal evaluation procedures (e.g., peer ratings, self-ratings) and, as well, to the informal word-of-mouth network. The same criteria regarding reliability and validity that are applied to student ratings should be equally applied to these other procedures. The issue, therefore, is not whether a formal evaluation procedure is reliable and valid, but rather whether a procedure, formal or informal, meets an acceptable criterion of reliability and validity in comparison with other procedures currently operating.

The present paper deals with a validity issue frequently used to criticize student ratings. It is based on the belief that being an entertainer and telling good jokes are sufficient for an instructor

to obtain favourable ratings from students, even in the absence of adequate course material. If student ratings assess effective teaching, and if effective teaching implies student achievement, then instructors who present inadequate content in their lectures should be rated unfavourably. A charismatic instructor can persuade students to believe that they are receiving quality instruction, when in fact they are learning very little. To the extent that ratings of a charismatic instructor do not correspond to content presented, some critics argue that the ratings are invalid.

B. Validity: Accuracy of Student Ratings

Student ratings are intended to provide an accurate assessment of an instructor's teaching effectiveness. Accuracy of measurement in questionnaire design is referred to as validity, that is, the ability of an instrument to assess what it is purported to be measuring. Thus, a weighing scale must measure a person's weight and not the person's height. An intelligence test must measure I.Q. to be valid and not personality, emotional stability, or social skills. Similarly, a student ratings questionnaire must measure teaching ability and not other unrelated factors associated with the classroom setting.

Standard test validation procedures can involve several kinds of validity such as face validity, predictive validity, concurrent validity, etc. Developers and users of student ratings questionnaires frequently emphasize construct validity, that is, the degree to which a variety of variables based on a common set of assumptions, correspond to each other. Thus, for example, a set of assumptions related to teaching effectiveness may include student satisfaction and student learning.

Objective, measurable criteria associated with these two concepts such as student ratings of satisfaction with the learning experience, and student achievement would be compared. The degree to which these two variables correlate with each other would represent a validity estimate.

Numerous criteria have been used to validate student ratings. Teacher-related criteria include faculty ratings, self-ratings, research productivity, education (e.g., years of teaching, number of courses). Student ratings validated using faculty ratings would suggest that, for a given instructor, student and faculty ratings correspond. That is, instructors rated favourably by students should also be rated favourably by faculty, and similarly, instructors rated unfavourably by students should also be rated unfavourably by faculty. Student-related criteria include achievement, current enrollments, continuation in subsequent courses, success in future careers, etc. Costin, Greenough, and Menges (1971) provide an excellent review of the validation research using many of these criteria.

One of the most widely-used criteria for validating student ratings appears to be student achievement as measured by multiple-choice exams, final grades, etc. A positive correlation would indicate that instructors who receive favourable ratings from students also tend to produce high achievement, and instructors who receive unfavourable ratings produce low achievement. Researchers have studied the ratings/achievement relationship in both the laboratory and the field setting. The field study does not require generalization of results to the natural setting, however, it lacks the precision and control of laboratory experimentation. Typically, a strong field study design would involve a multisection course, taught by

the numerous instructors, having common teaching objectives, syllabus, text book, testing procedures, etc. Initial differences in student characteristics between classes, especially student ability and knowledge of course topics due to lack of random assignment are taken into account in calculating final achievement outcomes. This design permits variables influencing achievement other than the instructors to be controlled, such that student achievement is teacher-produced. An instructor's impact on learning is the mean achievement score for the class. The ratings/achievement relationship is assessed by computing for each class a mean student ratings score and a mean student achievement score, then calculating a correlation coefficient based on these class means. The student ratings are considered valid if a significant, positive correlation is found.

Beginning with Remmers (1950), numerous researchers have conducted field studies to examine the ratings/achievement relationship, but few have incorporated the basics of proper experimental design. Rodin and Rodin (1972) provide a good example of a poorly-designed field study. In a large under-graduate calculus course, students met for lectures delivered by an instructor three times a week, and met for tutorials conducted by graduate teaching assistants twice a week. The instructor was responsible for organizing and presenting the course material and setting the tests, while the teaching assistants set tutorial problems and answered questions about the lecturers. Students were permitted to change tutorials between semesters if they wished. At the end of the course students were asked to rate their teaching assistants on the following item: "What grade would you assign to his total teaching performance? (A,B,C,D,F)." Achievement was measured using forty paradigm problems

which the students could retake up to six times. Rodin and Rodin found a -.75 correlation between student ratings and achievement, and concluded that "Students rate most highly instructors from whom they learn least" (P. 1164).

Several methodological problems make Rodin and Rodin's results highly questionable. First, random assignment was not used to allocate students or teaching assistants to tutorials. In fact, students were permitted to change tutorials, which was likely to produce major differences in their composition. A statistical procedure used to adjust achievement outcomes to account for initial differences in student ability actually biased the results against finding a positive correlation. Second, the item assessing teaching effectiveness was atypical in its construction and its focus. As a single-item measure, it departs from standard procedures which usually include several items. Those that do consist of only one item typically use verbal descriptions of each of the rating categories. Also, the focus is on the teaching assistant responsible for ancilliary tasks, rather than on the instructor who was primarily responsible for teaching and testing course content. Third, the achievement measure was based on final grades which are not as precise as test scores. This problem was compounded since there appears to be a lack of variability in the average grade received by the classes, with grades ranging from 3.0 to just over 3.5. Finally, the correlation coefficient is based on eleven observations. As Doyle (1975) points out, the confidence interval for this statistic is very wide and could easily range between $\pm$.60,

making its significance suspect. In sum the methodological weaknesses of this study are substantial and do not provide any useful evidence regarding the validity of student ratings.

Although better designed field research has found more encouraging results, it has not received the same acclaim. For instance, Sullivan and Skanes (1974) randomly assigned approximately 2300 students and over 100 instructors to multiple sections of introductory biology, chemistry, mathematics, physics, and psychology courses. Students rated their instructors' teaching on a single-item measure and wrote an essay-type final examination that was marked by a committee. Correlations were computed by comparing mean student ratings and achievement for each section in the five courses. The overall correlation computed across 130 sections was +.39, and for individual courses the correlations ranged between -.28 to +.57, with all but two being greater than +.32.

Instructor experience was found to have a significant impact on the magnitude of the correlations with magnitudes increasing with years of experience. For example, the ratings/achievement correlations for psychology instructors were +.01 for inexperienced faculty and part-time teaching assistants, +.13 for faculty having several years experience (e.g., associate professors), and +.69 for faculty having many years experience (e.g., full professors). Sullivan and Skanes interpreted these results as differences in teaching objectives, with experienced instructors being primarily concerned with student achievement and inexperienced instructors more concerned with having good rapport with the students and stimulating their interest. Accordingly, the experienced instructors' objectives are more

likely to be reflected in better final examination performance. This analysis may provide a partial explanation for Rodin and Rodin's results since they used graduate student teaching assistants. Subsequent research using random assignment of students to classes by Centra (1977) and Ellis and Rickard (1977) have also reported moderate positive ratings/achievement correlations.

The field validation research has been inconsistent in finding a relationship between student ratings and achievement (e.g. Feldman, 1976; Kulik & McKeachie, 1975), however, with correlations ranging from -.80 to +.72. This inability to find consistent, positive correlations is frequently interpreted by critics as supporting the invalidity of student ratings. But the discrepancy appears to be due to weak experimental designs (e.g., Rodin & Rodin, 1972) which lack random assignment, proper units of analysis, uniformity of course material, standardized achievement test properly-developed ratings questionnaires, etc. (Abrami, Leventhal, & Perry, 1977; Leventhal, Abrami, & Perry, 1977). Better designed field research appears to support consistent, positive ratings/achievement correlations (e.g., Centra, 1977, Ellis & Rickard, 1977; Frey, 1973; McKeachie, Lin, & Mann, 1971; Sullivan & Skanes, 1974).

Laboratory researchers have focused on how various teaching behaviors affect different student outcomes including student ratings and achievement. They appear more concerned with the systematic manipulation of teaching behaviors and their impact on students, rather than with the ratings/achievement relationship. Often the experiments take place in well-controlled laboratory settings, involving simulated classroom conditions

in which select teaching behaviors are manipulated, and using precise outcome measures. Advantages gained from the utilization of optimal experimental design procedures are compromised by problems with the generalizability of results to the natural classroom setting. Nevertheless, this literature does provide further evidence regarding the validity of student ratings. Much of it is based on the premise that effective teaching involves student learning, and that student ratings provide an estimate of teaching effectiveness. If, under some experimental conditions, student ratings and achievement show great discrepancies, then this can be used to question the validity of student ratings.

Educational seduction represents a series of studies having this orientation. The remaining sections of this paper provide a detailed review of this research. Section II introduces the concept of educational seduction, its relationship to the validity issue, and reviews the empirical results. Section III presents some implications of this research for the evaluation and improvement of teaching, and for future directions in research on teaching.

## II. EDUCATIONAL SEDUCTION: THE DR. FOX EFFECT

### A. Introduction

The concept of educational seduction comes from a widely-held belief in some academic circles that an entertaining, charismatic instructor can receive favourable student ratings while offering insufficient lecture content. Accordingly, critics claim that student ratings are invalid. This conclusion appears to be based on a model of teaching effectiveness which suggests that effective teaching involves student learning. For

student ratings to be valid, they must accurately reflect course content. Favourable ratings given under insufficient content conditions do not correspond to the poor quality of instruction, and therefore, student ratings are invalid.

Researchers have attempted to translate this belief into a testable hypothesis to determine whether it has empirical support. Naftulin, Ware, and Donnelly (1973) first introduced the terms "educational seduction" and the "Dr. Fox effect". They speculated that "Given a sufficiently impressive lecture paradigm, audience numbers at a lecture can feel satisfied that they have learned despite irrelevant, conflicting, and meaningless content conveyed by a lecturer" (P. 631). They trained a Hollywood actor to present material using double talk, contradictory remarks, and nonsequiturs, coupled with parenthetical humour and irrelevant examples. At a national conference on continuing education, the actor was introduced using an impressive curriculum vitae as Dr. Myron L. Fox, an expert on the application of mathematics to human behavior. He gave a one-hour lecture, followed by a half-hour question period, to a group of psychologists, psychiatrists, and social workers on a fictitious topic entitled "Mathematical Game Theory as Applied to Physical Education". The presentation was designed to be enthusiastic and entertaining, but substantially devoid of meaning and content. It was also videotaped and presented to a group of medical educators and to a group of administrators in a graduate education course. Following the presentation the groups completed a seven-item, forced-choice (yes/no) questionnaire.

The educators reported that the presentation was interesting and that it stimulated their thinking. Some requested more information about

the topic, and one person claimed to have read one of the nonexistent references. The educators appeared quite prepared to give favourable ratings based on an instructor's lecturing style without careful consideration to what was being taught. Naftulin et al. concluded than an entertaining, charismatic instructor can create the illusion of a positive learning experience, even in the absence of adequate lecture material. They suggested that their results support the argument that student ratings are invalid measures of teaching effectiveness because the ratings do not reflect the amount of actual learning taking place.

Beginning with Gage (1974) and Kaplan (1974), researchers have criticized Naftulin et al.'s conclusions on conceptual, methodological, and statistical grounds. Four methodological problems have repeatedly been mentioned. First, considering the conclusions Naftulin et al. wanted to draw, they failed to provide appropriate comparison groups as part of their experimental design. To test whether an entertaining instructor, who given inadequate lecture material, can produce favourable ratings requires at least one other lecture condition, i.e., either an entertaining instructor who provides <u>adequate</u> content, or a <u>nonentertaining</u> instructor who provides inadequate content. Ratings can then be compared between the original lecture condition and either of the two additional ones. Comparing the entertaining inadequate content condition with the entertaining, adequate content condition would provide the proper test of Naftulin et al.'s hypothesis. That is, the entertaining instructor who provides inadequate content should obtain comparable ratings to the entertaining instructor who provides adequate content. The effect of entertainment could also be tested by comparing the entertaining, inadequate

content condition with the nonentertaining, inadequate condition. Naftulin et al.'s design permits comparisons to be made only between their three educator groups which were exposed to the same lecture condition (entertaining, inadequate content). Valid conclusions, therefore, can only refer to rating differences caused by characteristics of the educators such as amount of education, profession, age, etc.

Second, Naftulin et al. used an atypical questionnaire which departed significantly from standard questionnaires (e.g., Hildebrand & Wilson, 1971; McKeachie, Lin, & Mann, 1971) in format and computation procedures. The questionnaire consisted of seven forced-choice items (yes/no), and the data analysis involved a simple summation of "yes" and "no" responses. More "yes" than "no" responses were interpreted as support for their hypothesis. This procedure is particularly suspect in view of the fact that a "yes" response to at least one item, "the instructor dwelled upon the obvious" should not be regarded as a favourable response. Most student ratings questionnaires use scales having five or more categories in which means are computed when more than one item is involved. To generalize their results to commonly used student ratings questionnaires would require the use of a more standardized format.

Third, Naftulin et al. did not assess student learning. Support for their hypothesis requires some estimate of student learning to compare to the student ratings. Presumably favourable student ratings coupled with very poor student learning would confirm their hypothesis. Student learning can be assessed by objective achievement tests that are common to most classroom settings. Subjective estimates of the amount learned can also be used, but are less precise. Naftulin et al.'s subjects did

not write an achievement test, nor were they even asked to rate how much they learned. The simple act of writing a test, or judging how much was learned may be sufficient to alert the subject to the inadequacies of the presentation. This activity may, in turn, result in subjects revising their initial opinion of the instructor to a less favourable position.

Finally, the length of the presentation was short and does not represent the extensive contact that students have with their instructor in the natural classroom setting. During a three month period students have numerous opportunities to interact with their instructor and to consult their notes and text books for additional information. In these conditions, students are more likely to recognize inadequate lecture presentations and revise initially favourable opinions. A proper test of Naftulin et al.'s hypothesis would require repeated presentations comparable to the classroom setting before students make their ratings.

Reed Williams and John Ware responded to these criticisms by conducting a series of experiments at Southern Illinois University. Their research represents the first systematic attempt to study educational seduction. Major modifications to Naftulin et al.'s original study included a better experimental design, a standard student ratings questionnaire, and a student achievement test. Incorporating these modifications, Ware and Williams (1975) reported evidence compatible with an educational seduction interpretation which was supported by Williams and Ware (1976) under different student incentive conditions, and by Williams and Ware (1977) after repeated presentations. Throughout these studies, Williams and Ware maintain that their data demonstrate that an entertaining (expressive) instructor can obtain favourable ratings in the absence of

adequate lecture content. They concluded, therefore, that their data supports the argument that student ratings are invalid.

Despite these modifications other researchers claim existing data does not indicate that student ratings are invalid. In studying four simulated classroom conditions, Perry, Abrami, and Leventhal (1979) did not find support for student ratings invalidity in three of the four conditions, including the condition most similar to Williams and Ware's experimental manipulations. They concluded that the effect is not readily replicated, and it does not appear to generalize to other classroom conditions. Meier and Feldhusen (1979) investigated educational seduction in relation to how student ratings are used: for diagnostic purposes to improve the quality of instruction, or for administrative purposes to make tenure, promotion, and salary decisions. Their results were not consistent with an educational seduction interpretation. Finally, Ramagli and Greenwood (1981) examined the effects of varying levels of instructor entertainment and lecture content. Their results were opposite to Williams and Ware's in that students were sensitive to content differences for entertaining instructors, but insensitive for nonentertaining instructors.

B. Experimental Paradigm

Educational seduction refers to an entertaining instructor who obtains favourable ratings while presenting insufficient course content. As discussed in Section IIA, testing this hypothesis requires an appropriate experimental design involving two independent variables: instructor expressiveness and lecture content. Instructor expressiveness, based on Naftulin et al.'s reference to seduction, charisma, and

entertainment, has evolved over the course of the research. Ware and Williams (1975) used the term "seduction", operationally defining it as" ... enthusiasm, humor, friendliness, expressiveness, charisma, personality" (P. 151). Williams and Ware (1976) reconceptualized "seduction" as expressiveness, and defined it as "... enthusiasm, humor, friendliness, charisma, and personality" (P. 50). Both Meier and Feldhusen (1979) and Ramagli and Greenwood (1981) used Williams and Ware's videotapes, and adopted their definition of expressiveness.

Perry, Abrami, and Leventhal (1979) and Perry, Abrami, Leventhal, and Check (1979) constructed their own videotapes, and developed a behavioral definition of expressivenss. Their primary concern with Williams and Ware's definition was its lack of precision. Defining expressiveness as enthusiasm, friendliness, personality, etc. provides synonomous terminology, but does not describe the specific teaching behaviors which make it up. Perry et al. defined expressiveness according to four constituent teaching behaviors: physical movement, voice inflection, eye contact, and humour. The frequency of occurrence of these constituent behaviors determined the different levels of expressiveness: low frequency defined low expressiveness, and high frequency defined high expressiveness. A psychology professor was provided with a role-sketch of instructor expressiveness and coached in role-playing the constituent teaching behaviors. Judges trained in teaching effectiveness rated the presentation on a ten-point scale. Only those presentations having a mean of eight or greater were retained on videotape. Students rated the low and high expressiveness videotapes on the following dimensions: enthusiasm, humour, friendliness, physical movement, vocal inflection, personality style, warmth, intelligence,

pragmatic character, and dynamic character. All multiple t comparisons between the low and high expressive conditions were significant ($p<.001$), with high expressiveness always evaluated more favourably.

Lecture content refers to the amount of material contained in the presentation. Ware and Williams (1975) manipulated lecture content by varying the number of teaching points covered during the lecture. They had three levels of lecture content, low, medium, and high, which was defined as 4, 14, and 26 teaching points. To maintain a comparable presentation length of approximately 30 minutes, teaching points eliminated from the low and medium content conditions were replaced with filler involving: discussions of unrelated examples, events, etc.; repeated emphasis on what was to be covered without such coverage; circular discussions of unrelated concepts. Perry et al. (1979) used a similar procedure to manipulate lecture content with 12 and 24 teaching points defining the low and high content conditions.

Two dependent variables are used to measure the effects of instruction: student ratings and student achievement. Student ratings measures have included diagnostic multiple-item questionnaires such as McKeachie's scale (McKeachie, Lin, & Mann, 1971), and Pohlmann's scale (Pohlmann, 1975), and a single-item summative measure that was validated using student achievement (Sullivan & Skanes, 1974). Ideally, a standardized student ratings measure is preferred which is validated using some other criterion measure such as student achievement. Generalizability of the results to the field setting is enhanced if widely-accepted measures are used based on proper psychometric development. Student achievement is assessed using multiple-choice tests based on the high content lectures

and designed to measure factual knowledge and comprehension. Williams and Ware and Perry et al. disagree about the significance of the achievement test. Ware and Williams (1979; 1980) argue that the achievement test is intended primarily as a manipulation check on lecture content. Proper manipulation of lecture content means that student achievement will vary significantly between conditions. Perry et al. consider achievement to be an integral part of the instruction process and their analysis of educational seduction. Both these viewpoints are developed in greater detail in the next section (IIC).

Educational seduction can be examined by combining the two independent variables, expressiveness and content, and the two dependent variables, student ratings and achievement, into an experimental design. The basic design is a two by two factorial, having two levels of expressiveness (low, high), and two levels of lecture content (low, high). Table 1 represents the basic design and provides some hypothetical data. For purposes of this discussion, the dependent variables have been simplified such that both student ratings (SR) and student achievement (SA) have a range from zero to ten. Each cell in the two by two matrix has two scores: the upper left-hand score is the student rating mean, and the lower right-hand score is the student achievement mean. Dr. Fox is represented by the high expressive low content cell in which student ratings do not correspond with achievement. Naftulin et al.'s original study reproduced only this one cell of the two by two matrix.

These results may be interpreted by some as supporting an educational seduction hypothesis. For the low expressive instructor student ratings appear to correspond to the amount of lecture content.

TABLE 1: Paradigm depicting a hypothetical set of results that may be used to argue the invalidity of student ratings.

| | | LECTURE CONTENT | | | |
|---|---|---|---|---|---|
| | | Low Content | | High Content | |
| | | SR | SA | SR | SA |
| INSTRUCTOR EXPRESSIVENESS | Low Expressiveness | 2 | 2 | 5 | 5 |
| | High Expressiveness | 8 | 5 | 8 | 8 |

SR = Student ratings

SA = Student achievement

As lecture content increases, ratings become more favourable. Furthermore, student achievement is compatible with lecture content, with student performance improving as lecture content increases. Consequently, it may be argued that student ratings are valid since they correspond to lecture content and to student achievement. For the high expressive instructor, student ratings do not correspond with the amount of lecture content. As lecture content increases, ratings remain unchanged. Student achievement however, is consistent with lecture content and reveals a similar pattern to low expressiveness conditions. It is argued, therefore, that student ratings are invalid since they do not correspond to lecture content or to student achievement. It should be noted that the hypothetical data in Table 1 represents one possible set of results, and is not intended to be definitive. The research literature suggests that the empirical results are much more complex.

C. Hypothesis-Testing

An analysis of Table 1 suggests that there may be several approaches to educational seduction. Each approach should specify the kinds of statistical comparisons and the set of results that are required to demonstrate the invalidity of student ratings. At least two approaches appear to have evolved in the literature based on somewhat different models of teaching effectiveness and validity: Williams and Ware, and Perry, Abrami, and Leventhal. The following discussion reviews these two approaches in detail. It should be emphasized that what is referred to as Williams and Ware's approach is inferred from their research since it has not been clearly articulated. Attempts to clarify their approach (Ware & Williams,

1971; 1980) has resulted in further criticisms by Frey (1979; 1980) and Leventhal (1979; 1980).

1. Williams and Ware. Williams and Ware base their approach to educational seduction on a process model of instruction. According to this approach Ware and Williams (1979) consider the validity of student ratings "... in relation to experimentally manipulated differences in teaching process (i.e. content coverage and expressiveness) in lecture preparations" (P. 7). Their focus is on student ratings in relation to lecture content. Student ratings are invalid if they do not correspond to the amount of material presented during the lecture. Student achievement is considered to be a manipulation check of the content variable. That is, content is considered manipulated properly only if significant achievement differences are found between content conditions.

Williams and Ware's approach suggests that several statistical comparisons and a specific pattern of results are necessary to test for educational seduction. First, student achievement must be compared between content conditions for each expressiveness level. Successful manipulation of the content variable is indicated by significant achievement differences between content conditions. Second, student ratings must be compared between expressiveness conditions to determine whether it has been properly manipulated. Third, student ratings for the low expressive instructor must be compared between content conditions to determine whether ratings become more favourable with increasing lecture content. Fourth, student ratings for the high expressive instructor must be compared between content conditions to determine whether ratings remain unchanged.

It should be emphasized that, while Williams and Ware's approach focuses primarily on the third and fourth sets of comparisons, the first and second sets are equally important. For example, suppose that all the achievement means in each cell in Table 1 were similar ($\bar{X}s=5$). Although the ratings results are compatible with Williams and Ware's approach, absence of achievement differences between content conditions (set one comparisons) would seriously jeopardize any meaningful interpretation of the ratings results based on the third and fourth sets of comparisons.

Analysis of variance procedures can be used in conjunction with multiple comparison techniques to perform these statistical comparisons. The first and second sets of comparisons can be determined from main effects of content on achievement and expressiveness on ratings. However, testing main effects does not provide exact information regarding all comparisons when more than two levels of the independent variables are involved, nor does it provide information regarding comparisons for each expressiveness level. Under these conditions a priori multiple comparisons are preferred. The third and fourth sets of comparisons are based on an Expressiveness by Content interaction on student ratings. Multiple comparisons should indicate a specific form of this interaction. That is, significant rating differences should occur between content conditions for the low expressive instructor coupled with insignificant differences between content conditions for the high expressive instructor. Other forms of this interaction would not be compatible with an educational seduction interpretation, as for example, rating differences between content conditions for the high expressive instructor coupled with no differences between content conditions for the low expressive instructor (e.g., Ramagli & Greenwood, 1981), or

rating differences between expressiveness conditons for one content condition, but not the other.

Williams and Ware have been inconsistent in performing these statistical comparisons and in obtaining a pattern of results necessary to support an educational seduction interpretation. Ware and Williams (1975) used the factorial design described in Table 1 with one modification, lecture content had three levels instead of two. They concluded from their results that educational seduction was supported, however, not all statistical comparisons were performed, nor was the obtained pattern of results compatible with the predicted pattern. As part of the first set of comparisons, they did not test content effects on achievement for both the low and high expressive instructors. Inconsistencies in obtained results were: (1) achievement differences between content conditions were absent for the low and medium, and low and high comparisons; (2) ratings differences between expressiveness conditions were not found for the high content conditions, and (3) for the low expressive instructor, ratings differences did not occur for the low and medium, and the medium and high content comparisons. Williams and Ware (1976) employed the Expressiveness (low, high) by Lecture Content (low, medium, high) factorial with the addition of a student incentive variable (no incentive, incentive). They did not perform the four sets of comparisons described above for each incentive level, but rather combined the results across incentive levels. Again, they did not test content effects on achievement for both the low and high expressive instructors. Inconsistencies in obtained results were: (1) ratings differences between expressive conditons were not found for the medium and high content

conditions, and (2) for the low expressive instructor, ratings differences did not occur for the medium and high content condition. Finally, Williams and Ware (1977) used the same Expressiveness by Content factorial in conjunction with a repeated presentation variable (presentation one, two). They did not analyze their results separately for each presentation level, but rather tested the combined results. Once again, not all the required statistical comparisons were performed, nor was the pattern of results congruent with the predicted pattern. They did not test content effects on achievement for each expressiveness level, nor did they test expressiveness effects on ratings for each content level. Inconsistencies in obtained results occured in that: (1) achievement differences between content conditions were absent for the low and medium, and the low and high comparisons; (2) ratings differences between expressiveness conditions were not reported, and (3) for the low expressive instructor, ratings differences did not occur for the low and medium content comparison.

Meier and Feldhusen (1979) investigated educational seduction using Williams and Ware's approach and their videotapes. They adopted a modified Expressiveness by Content factorial design in which lecture content had two levels, medium and high. They added a third variable, purpose of evaluation, informing students that the ratings would be used either for administrative decisions regarding tenure, promotion, and merit, or for feedback to improve the quality of instruction. The four sets of comparisons were not performed for each purpose of evaluation condition, but rather on the combined data for both purpose conditions. Inconsistencies in obtained results were: (1) no achievement differences were found between content conditions; (2) no ratings comparisons between low and

high expressiveness were reported for each content condition and (3) no Expressiveness by Content interaction was found, suggesting that content effects on ratings were comparable for both the low and high expressive instructor. Finally, Ramagli and Greenwood (1981) used an Expressiveness (low, high) by Content (low, high) design combined with a second Expressiveness by Content factorial design. Again, the four sets of comparisons were not done. Inconsistencies in obtained results were: (1) no achievement comparisons were made; (2) no ratings comparisons between low and high expressiveness were reported for each content condition, and (3) the expressiveness by content interaction was opposite to predicted results in that, for the low expressive instructor, content did not affect ratings, but for the high expressive instructor, content did affect ratings.

Table 2 presents the statistical comparisons and the predicted pattern of results necessary for an educational seduction interpretation. It also summarizes the obtained results for Williams and Ware's studies, and for the others based on their approach. The word "NO" in a cell indicates that no comparisons were performed, nor results reported. The word "YES" indicates that the required comparisons were done, and the obtained pattern of results was consistent with predicted EDUCATIONAL SEDUCTION pattern of results. Cells without "NO" or "YES" specify obtained results which were consistent (not underlined) and inconsistent (underlined) with the predicted EDUCATIONAL SEDUCTION results. Support for educational seduction would require a combination of four "YES" outcomes across the table for a particular study.

It is clear from Table 2 that support for educational seduction is

TABLE 2: Comparison of predicted results with obtained results for Williams and Ware's approach to Educational seduction.

| | STATISTICAL COMPARISONS | | | |
|---|---|---|---|---|
| | Set 1 | Set 2 | Set 3 | Set 4 |
| | Student achievement compared between content conditions for each expressiveness level | Student ratings compared between expressiveness conditions for each content level | For low expressive instructor, comparison of ratings between content conditions | For high expressive instructor, comparison of ratings between content conditions |
| EDUCATIONAL SEDUCTION OUTCOMES (PREDICTED) | LC<MC<HC for LE<br>LE<MC<HC for HE | LE<HE for LC<br>LE<HE for MC<br>LE<HE for HC | LC<MC<HC | LC=MC=HC |
| Naftulin, Ware & Donnelly (1973) | NO | NO | NO | NO |
| Ware & Williams (1975) | MC<HC<br>LC<HC<br>MC=HC<br>No content comparisons made for each expressiveness level | LE<HE for LC<br>LE<HE for MC<br>LE=HE for HC | LC<HC<br>LC<MC<br>MC=HC | YES |
| Williams & Ware (1976) | LC<MC<HC<br>No content comparisons made for each expressiveness level | LE<HE for LC<br>LE=HE for MC<br>LE=HE for HC | LC<HC<br>LC<MC<br>MC=HC | YES |
| Williams & Ware (1977) | MC<HC<br>LC=MC<br>LC=HC<br>No content comparisons made for each expressiveness level | NO | LC<HC<br>LC<MC<br>LC=MC | YES |
| Meier & Feldhusen (1979) | MC=HC | LE<HE<br>No expressiveness comparisons for each content level | MC=HC | NO |
| Ramagli & Greenwood (1981) | NO | LE<HE<br>No expressiveness comparisons for each content level | LC=HC | LC<HC |

LE = low expressiveness
HE = high expressiveness

LC = low content
MC = medium content
HC = high content

extremely weak. First, there does not appear to be one study which carried out the four sets of comparisons completely. This is most evident for the first set in which comparisons were made between content conditions, but not for each expressiveness level. Although seemingly inconsequential, this omission can be serious if content affects achievement for the low, but not the high, expressive instructor similar to its impact on student ratings. Again, refer to Table 1 and suppose that for the high expressive instructor, the mean achievement outcomes for both content conditions were similar ($\bar{X}$s=8). In this case, the third and fourth sets of comparisons involving only student ratings supports Williams and Ware's approach to educational seduction. However, including the first set of comparisons involving student achievement suggests a totally opposite conclusion: that student ratings are valid in terms of the ratings/achievement correlation. Second, there appears to be a lack of consistency in results for comparisons in common between studies. For example, Williams and Ware (1976) found content effects on student achievement for all comparisons, whereas Ware and Williams (1975) and Williams and Ware (1977) found effects for only the medium and high content comparison, while Meier and Feldhusen (1979) reported the same medium and high comparison insignificant.

It should be reiterated that, although Williams and Ware's approach focuses on the third and fourth sets of comparisons, it does not rule out the necessity for the first and second sets of comparisons. As discussed above, omitting these comparisons can lead to an opposite conclusion than if they are included. Even considering only the primary comparisons involved in Williams and Ware (sets three and four), the data do not support an educational seduction interpretation. For the low expressive

instructor, content effects on ratings are inconsistent between Williams and Ware's studies, and non-existent in Meier and Feldhusen's and Ramagli and Greenwood's studies. For the high expressive instructor, the insignificant effect of content on ratings was found only in Williams and Ware's studies, was unreported in Naftulin et al.'s and Meier and Feldhusen's studies, and was significant in Ramagli and Greenwood's study. These anomolies, combined with other results of the four sets of comparisons presented in Table 2, make a definitive statement supporting educationable seduction untenable.

2. Perry, Abrami, and Leventhal. Perry, Abrami, and Leventhal (1979) consider educational seduction to be not just a validity issue, but also an opportunity to study causal relationships involved in the teaching process. They argue against the over-emphasis on such terms as "educational seduction" and the "Dr. Fox effect", since they divert attention away from the more important issue of research on teaching. As a validity issue, they postulate that educational seduction involves both the expressiveness and content variables, and student ratings and achievement. For a low expressive instructor, ratings and achievement correspond to content differences, and for a high expressive instructor, ratings do not and achievement does correspond to content differences (refer to Table 1). This relationship suggests a more complex form of the Expressiveness by Content interaction having two dependent variables. It involves an Expressiveness by Content interaction on student ratings and no interaction effect on student achievement.

This complex interaction requires a statistical procedure which determines whether content affects ratings and achievement differently for each expressiveness conditon. Gabriel and Hopkins (1974) discuss three

analysis of variance (ANOVA) procedures for testing interaction effects involving multiple dependent measures. The first procedure requires a separate univariate ANOVA for each dependent variable which they recommend against because: (a) the overall probability for Type 1 errors for the set of significance tests cannot be assessed, and (b) there is no method for testing whether a particular variable affects one dependent measure significantly different from another dependent measure. A second procedure requires converting the multiple measures to a common metric unit and performing a multivariate ANOVA (MANOVA). A third procedure involves a repeated measures Anova in which the dependent measures are also converted to common metrics and analyzed as a repeated measures variable. For this procedure, the multiple dependent measures can be converted to $z$ scores such that each measure becomes a separate level of an independent variable called measures. Accordingly, the measures variable can be combined with the other independent variables in the experimental design and analyzed using an univariate repeated measures ANOVA.

The second and third procedures are both acceptable, although the repeated measures ANOVA has greater power for significance tests involving the repeated measures factor. The increased power results from covariance assumptions which are not violated when the repeated measures factor has only two levels (McCall & Appelbaum, 1973). Consequently, an advantage of the second and third procedures is that ratings and achievement can be compared directly. The repeated measures procedure is more appropriate for Perry, Abrami, and Leventhal's approach, since it permits interaction terms involving the measures variables and the other independent variables

to be tested directly. The interaction term of primary interest to their approach is the Expressiveness by Content by Measures interaction. This interaction tests whether expressiveness influences the degree to which content affects ratings differently than achievement. The raw ratings and achievement scores are transformed to common $z$ scores which are based on the sample mean and the variance of the measure. The variance for a measure is computed for each treatment group separately and pooled which excludes variation from group treatment effects (Gabriel & Hopkins, 1974).

Perry, Abrami, and Leventhal (1979) investigated educational seduction under varying classroom conditions. They combined the Expressiveness by Content factorial with two other classroom variables: opportunity to study (unavailable, available) and student incentive (low, high). The Expressiveness by Content by Measures interaction was tested under four different classroom conditions: no opportunity to study--low incentive; no opportunity to study--high incentive; opportunity to study--low incentive; opportunity to study--high incentive. Assessing the interaction effect under various levels of the independent variables is another procedure which differs from Williams and Ware's approach. Their results indicated that educational seduction as defined by the Expressiveness by Content by Measures interaction was found for only one classroom condition: no opportunity to study--high incentive. This classroom condition differs from previous educational seduction research because of the incentive variable, and from Williams and Ware (1976) because incentive was manipulated differently. A significant Expressiveness by Content by Measures interaction in the no opportunity to study--low incentive condition was required to replicate previous demonstrations of educational seduction. On balance, Perry, Abrami, and Leventhal

(1979) concluded that educational seduction was not replicated.

Perry, Abrami, and Leventhal's study (1979) is important here becasue it presents a systematic approach to educational seduction, and because it represents the most rigorous attempt to replicate Williams and Ware's research. Their inability to replicate suggests that educational seduction is not a reliable effect. Small differences between research laboratories in videotape format, lecture topics, instructors, students, etc. appear sufficient to prevent its replication. This is especially disconcerting since Perry et al.'s procedures had greater similarity to the natural classroom. It seems all the more unreasonable, therefore, to claim that educational seduction can be generalized from the laboratory to the field setting where classroom differences are even greater.

D. Conclusions

The education seduction research has attempted to determine whether there is empirical support for the belief that an entertaining instructor can receive favourable ratings while providing inadequate lecture content. The researchers have come to opposite conclusions. Williams and Ware argue that their research supports educational seduction, and that the effect can be generalized to classrooms in the field setting. Specifically, Ware and Williams (1979) come to the following conclusions:

> "... (a) differences in content coverage produce corresponding differences in student achievement, (b) student ratings scored elaborately (e.g., weighted combinations of items) or simply summed are very valid in relation to differences in lecturer enthusiasm, (c) student rating items (regardless of their content or how they are scored) do less well or fail to detect differences in lecturer content coverage, (d) the latter is particularly true when lectures differing in

> content coverage are delivered enthusiastically, and (e) findings leading to the preceding conclusions hold true with back-to-back exposures to lectures differing only in subject matter under the same conditions and also hold true under different incentive conditions.

Furthermore, they argue that Perry et al.'s failure to replicate their findings does not disconfirm the educational seduction effect. The focus on student achievement and the different experimental procedures do not make Perry et al.'s study a proper test of their approach to educational seduction. Other researchers adopting Williams and Ware's approach also report evidence in favour of educational seduction.

A closer examination of those same studies claiming support for educational seduction suggests a very different interpretation of the research. Table 2 provides an outline of the four sets of statistical comparisons implied by Williams and Ware's approach and necessary to test for educational seduction. It is clear from this analysis (see Section IIC) that, considering only those studies using Williams and Ware's approach, the four sets of comparisons were never performed in any one study, and major inconsistencies between studies occurred in comparisons done in common. Based on the analysis of those studies supporting educational seduction (Table 2), it does not appear that student ratings are invalid. This conclusion is opposite to Williams and Ware's position and is based on the same set of empirical results.

Perry, Abrami, and Leventhal (1979) came to a similar conclusion based on different experimental procedures, and on different empirical results. Their approach represents a more precise analysis to educational seduction than taken by those studies described in Table 2. Based on their failure to replicate, they concluded that an educational seduction effect is not

supported by empirical results. Furthermore, they urged extreme caution in generalizing any significant effects found in this type of research to classrooms in the field setting. Based on their own data, Perry et al. came to the same conclusion as the conclusion resulting from the Table 2 analysis: empirical evidence does not support educational seduction.

Abrami (1981) provides another interpretation of the educational seduction research. He identified ten studies to date which had used the Expressiveness by Content factorial design, and performed a meta-analysis employing omega-squared ($\omega^2$) as an estimate of effect size. Practical significant was considered as statistically significant effects which accounted for more than 5% of the variance. For each study effect on student ratings and achievement, size was calculated for expressiveness, content, and the Expressiveness by Content interaction. For all studies combined the expressiveness effect on ratings was large ($\bar{X}$=26.9%), but the content ($\bar{X}$=3.5%) and the Expressiveness by Content ($\bar{X}$=1.8%) effects were not of practical significance. The content effect on achievement was significant ($\bar{X}$=14.6%), but the expressiveness ($\bar{X}$=3%) and the Expressiveness by Content ($\bar{X}$=0.2%) effects did not reach practical significance. Based on his meta-analysis, Abrami concluded that instructor expressiveness has a major influence on student ratings but not on student achievement, and lecture content has a strong effect on achievement but not on ratings. It is clear from Abrami's data that the Expressiveness by Content interaction accounts for very little variance in student ratings or achievement. His data provide further evidence against educational seduction since the Expressiveness by Content interaction is a critical test of the hypothesis.

Taken together, the empirical results do not support educational seduction. Williams and Ware's claims to the contrary, the bulk of the evidence does not indicate that student ratings are invalid. The analysis presented in Table 2 of those studies claiming support for educational seduction suggests that the evidence is very weak. Perry, Abrami, and Leventhal (1979) provided additional data opposing an educational seduction interpretation. Finally, Abrami (1981) could not find sufficient evidence in a meta-analysis of educational seduction. Although existing research findings do not support the belief about the entertaining instructor, this does not mean that the belief is unfounded. It may be that the effect occurs only in certain classroom conditions which researchers have been unable to simulate. It may also be that the effect does exist, but because it is so weak it has very little impact in comparison to other classroom variables (see Abrami, 1981).

## III. IMPLICATIONS

### A. Evaluation of Teaching

The educational seduction research has at least two implications regarding the evaluation of teaching: common beliefs about validity, and administrative decision-making. First, the literature suggests that at least one common belief about student ratings validity may not be true. Periodically, strong opposition has developed against student ratings because of their supposed invalidity. Often, however, little consideration is given to the verocity of beliefs about student ratings invalidity, or to the validity of other evaluation systems such as the word-of-mouth network. The educational seduction research, which represents a systematic, empirical

analysis of one common belief, provides little evidence supporting such claims. It underscores the need for closer scrutiny of other common beliefs about the validity of student ratings. It also emphasizes the importance of a systematic approach to assessing our knowledge about student ratings.

Second, tenure, promotion, and merit decisions are often made in the absence of specific information about teaching. Consideration may be given to whether instructor A is a good or bad teacher, but not to what kinds of teaching behaviors instructor A possesses that are advantageous to the institution's goals and to student development. For example, consider the two high expressive instructors described in Table 1. Both produce favourable student ratings, but the high content instructor also produces more achievement. Some institutions may regard the combined effects of the high expressive/high content instructor as more advantageous to their goals, and wish to reward such an instructor accordingly. The educational seduction research draws attention to the effect that certain teaching behaviors have on students. Some institutions may wish to use a similar approach for more effective administrative decision-making.

B. Improvement of Teaching

Increasingly, more attention is being given to what Gage (1958) identified as one of the purposes of student ratings: the improvement of teaching. The educational seduction research provides some potentially useful information in this respect. It assesses the specific effects of certain teaching behaviors on students. This cause-effect appraoch can be useful for improving the general quality of instruction in an educational

institution. Instructors who wish to improve their teaching can obtain information on which behaviors are generally important, and what effects these behaviors have on students.

After deciding through consultation which behaviors are appropriate, they can set about to develop them based on the behavioral analysis provided by the research literature. Specifically, the literature describes the constituent components of the behavior, and its effects on students. For example, suppose instructor A is concerned about both student ratings and achievement, and wishes to improve his/her expressiveness. The research describes specific activities which must be developed such as physical movement, voice intonation, eye contact, and humour. It also indicates that lecture content must be taken into account to maximize student achievement. Instructor B, on the other hand, may only be concerned with student achievement and wish to focus on lecture content, by emphasizing specific teaching points. Furthermore, both instructors may wish to administer a teaching evaluation to determine whether their efforts have actually resulted in recognizable improvements.

C. Research on Teaching

Perhaps the most important contribution of the education seduction literature is its emphasis on experimental research. The experimental paradigm described in Table 1 offers a useful conceptual framework for research on teaching. The Expressiveness by Content factorial depicts a simulated instructor having only two teaching behaivors in his repertoire, and affecting only student ratings and achievement. Nevertheless, this two-dimensional simulation can be used as a basis to address a number of issues

related to effective teaching.

1. Effective Teaching. A good instructor possesses a number of teaching behaviors which students consider important for effective teaching. Numerous factor analytic studies have identified a finite number of these teaching behaviors such as skill, rapport, expressiveness, enthusiasm, etc. (e.g., Doyle, 1975; McKeachie & Kulik, 1975). Although the relationships between these teaching variables and student variables have been examined in correlational studies, systematic, cause-effect research is lacking. The Expressiveness by Content factorial presents a simulated instructor having only two of these teaching behaviors. It does, however, provide a framework within which to develop a multidimensional instructor based on the original two-dimensional instructor. Accordingly, the simulated instructor would be more representative of the instructors in the field setting. For example, if instructor skill and organization were identified as two effective teaching behaviors, they could be combined with the expressiveness and content by skill by organization factorial. Various combinations of these variables could be assessed to determine positive and negative effects on students. The enlarged design approximates more closely the multidimensional nature of instruction in the field setting, and it provides a useful analysis of how different teaching conditions affect students.

2. Instructor Reputation. Perry, Abrami, Leventhal, and Check (1979) investigated instructor reputation in relation to the Expressiveness by Content factorial. A student's initial classroom experiences may be mediated by expectations about the instructor, since a significant

number of students select classes based on instructor reputation (Leventhal, et al., 1975). Previous research (Perry, Niemi, & Jones, 1974) suggests that reputation-produced expectations may affect subsequent student ratings and possibly student achievement in accordance with initial expectations. Generally, instructor reputation involves certain beliefs, attitudes, and expectancies held by members of an academic community regarding an instructor's professional abilities and performances in such areas as teaching, research, and administration.

Perry et al. (1979) were interested in how instructor reputations interact with the teaching behaviors possessed by an instructor. They combined the Expressiveness by Content factorial with an instructor reputation variable, and focused on Instructor Reputation by Teaching Behavior interactions. Their results indicated that reputation interacted with expressiveness in that students rated the positive reputation/high expressive instructor more favourably than the negative/low expressive instructor. No reputation by content interaction was found. Their results suggest that instructor reputation interacts with some teaching behaviors, but not others. Moreover, only certain student variables are effected, i.e., student ratings but not achievement.

These results have implications for instruction in the field setting. The Reputation by Expressiveness interaction indicated that a negative reputation, high expressive instructor can receive lower student ratings than a positive reputation, high expressive instructor. Both high expressive instructors, however, may have teaching skills that contribute to and are valued by an educational institution. A negative reputation can result in these skills being underestimated in terms of student ratings.

Coincidently, low expressive instructors are unaffected by the reputation variable, although their contribution to the institution is less beneficial.

3. Learned Helplessness. Recently Perry and Leventhal (1980) have argued that Seligman's learned helplessness model can be applied to the classroom setting. The classroom can be viewed as a learning environment in which differential response/outcome contingency relationship operate. Consider the response/outcome contingency issue as a student characteristic. Certain student behaviors are necessary to achieve success in a course, such as class attendance, studying behavior, verbal fluency, questioning behavior, persistence at a task, etc. Absence of these behaviors will likely lead to failure. Thus, the occurence of a given response produces a specified outcome: taking notes and studying course materials will result in passing the course, while not taking notes and lack of studying will lead to failure. For both success and failure, the response/outcome relationships are dependent. Other classroom situations may involve response/outcome noncontingent relationships. Consequently, class attendance, note-taking, studying, etc., may not produce success consistently and their absence may not lead to failure consistently. Here the occurrence or absence of these responses are independent of success and/or failure. Under these circumstances the student may believe that s/he has very little control over the eventual outcome (pass/fail) in the course.

It is important to note that, although student failure can occur in both dependent and independent contingency conditions, it may have qualitatively different cognitive, motivational, and behavioral effects.

That is, the failing student who perceives the classroom setting as involving dependent relationships is likely to attribute failure to identifiable factors over which s/he has some control, including studying, note-taking, teaching effectiveness, etc. This student may attempt to reduce failure by modifying behaviors associated with those factors contributing to failure, e.g., studying more, taking notes, changing instructors, etc.; or by engaging in new behaviors which may facilitate success such as enrolling in a study methods course. The failing student who perceives the classroom as involving independent relationships is less likely to attribute failure to specific causes; or is likely to attribute failure to causes over which s/he has no control such as luck, intelligence, the educational system, etc. This student is unlikely to strive to reduce failure, and may be more prone to boredom, apathy, failure, absenteeism, and drop-out.

Consequently, within a given classroom students may differ according to previous contingency training experiences. Some students may have a training history of response/outcome dependence in which their behaviors have produced specific outcomes consistently, while other students may have a response/outcome independence training history in which their behaviors have not caused predictable outcomes. Thus, contingent-trained students may be more motivated, involved, and active due to expectations that they have some influence over classroom events. Noncontingent-trained students may be more bored, apathetic, and passive due to expectations that they have very little control over classroom developments.

Students may have different contingency training histories which permit them to be classified as contingent and noncontingent. An instructor's stimulus attributes consist, in part, of the teaching behaviors that s/he

exhibits in the classroom. Consequently, a contingency by teaching behaviors interaction may exist in that students with different contingency training histories come into contact with various classroom teaching behaviors. Accordingly, several research questions can be considered regarding this interaction: To what extent do teacher behaviors interact with student contingency training to affect student achievement, apathy, drop-out etc. Are there teacher behaviors which increase the deleterious effects of noncontingency training? Do certain behaviors remediate noncontingency training deficits?

This interaction can be represented in its simplest form by a two by two factorial design involving student contingency training (contingent, noncontingent) and a specified teaching behavior, e.g., instructor expressiveness (low, high). Various student performance measures can be used to assess effects such as student achievement, student ratings, attributions, etc. Thus, contingent-and noncontingent-trained students can be compared under different teaching conditions. A contingent-trained student may relate to a high expressive instructor differently than a noncontingent-trained student relates. Or, a noncontingent-trained student may respond differently to a high expressive teacher than to a low expressive teacher. More complex versions of this interaction would involve: other levels of contingency training and instructor expressiveness; other student characteristics, e.g., incentive, anxiety, and other teaching behaviors, such as lecture content, skill, organization, etc. The significance of this interaction is that it can test whether some teaching behaviors increase the deleterious effects of noncontingency training, and whether others may actually remediate noncontingency training effects.

## Footnotes

[1]This article is based on several studies conducted at the University of Manitoba. Support for these studies was obtained in part from the University of Manitoba research board, the Canada Council (S74-1078, S76-0345), and the Social Sciences and Humanities Research Council of Canada. Requests for reprints should be addressed to Raymond P. Perry, Department of Psychology, University of Manitoba, Winnipeg, Manitoba, Canada, R3T 2N2.

[2]The author would like to acknowledge the substantial contributions of Les Leventhal, Philip C. Abrami, and Wenda Dickens to the University of Manitoba research programme on the Dr. Fox effect. The viewpoints expressed in this article are those of the author, however, and do not represent the Manitoba research group as a whole.

## References


Abrami, P. C. Educational seduction. Unpublished manuscript. Concordia University, Montreal, Canada, 1981.

Abrami, P. C., Leventhal, L., & Perry, R. P. The generalizability of student ratings of instruction. Presented at the American Educational Research Association annual meeting, New York, April, 1977. (ERIC Document Reproduction Service No. ED 139 832).

Centra, J. Student ratings of instruction and their relationship to student learning. American Educational Research Journal, 1977, 14, 17-24.

Costin, F., Greenough, W. T., & Menges, R. J. Student ratings of college teaching: Reliability, validity, and usefulness. Review of Educational Research ,1971, 41, 511-535.

Doyle, K. O. Student evaluation of instruction. Lexington, Mass.: D. C. Heath and Co., 1975.

Ellis, N. R., & Richard, H. C. Evaluating the teaching of introductory psychology. Teaching of Psychology, 1977, 4, 128-132.

Frey, P. W. Student ratings of teaching: Validity of several rating factors. Science, 1973, 182, 83-85.

Frey, P. W. A two-dimensional analysis of student ratings of instruction. Research in Higher Education, 1978, 83-85.

Gabriel, R. M., & Hopkins, K. D. Relative merits of MANOVA, repeated measures ANOVA, and univariate ANOVAs for research utilizing multiple-choice criterion measures. Journal of Special Education, 1974, 8, 377-389.

Gage, N. L. Ends and means in appraising college teaching. Conference on Appraisal of Teaching in Large Universities. University of Michigan, Ann Arbor, 1958.

Gage, N. L. Students' ratings of college teaching: Their justification and proper use. In N. S. Glasman and B. R. Killait (Eds.) Second UCSB conference on effective teaching. Graduate School of Education and Office of Instructional Development, University of California, Santa Barbara, CA., 1974. Pp. 72-86.

Guthrie, E. R. The evaluation of teaching: A progress report. Seattle: University of Washington, 1954.

Hildebrand, M., & Wilson, R. Effective university teaching and its evaluation. Washington, D. C.: U. S. Department of Health, Education, and Welfare, Office of Education, 1970.

Kaplan, R. M. Reflections on the Dr. Fox Paradigm. Journal of Medical Education, 1974, 1974, 49, 310-312.

Leventhal, L. The Dr. Fox effect: An alternative interpretation. Instructional Evaluation, 1979, 4(1), 1-6.

Leventhal, L. Alternative interpretation of the Dr. Fox effect: Reply to Frey. Instructional Evaluation, 1980, 4(2), 13-14.

Leventhal, L., Abrami, P. C., & Perry, R. P. Bogus evidence for the validity of student ratings. Presented at the American Psychological Association Annual meeting, San Francisco, August, 1977. (ERIC Document Reproduction Service No. ED 150 510).

Leventhal, L., Abrami, P. C., Perry, R. P., & Breen, L. J. Section-selection in multi-section courses: Implications for the validation and use of teacher rating forms. Educational and Psychological Measurement, 1975, 35, 885-895.

McCall, R., & Appelbaum, M. Bias in the analysis of repeated-measures designs: Some Alternative approaches. Child Development, 1973, 44, 401-415.

McKeachie, W. J., Lin, Y.; & Mann, W. Student ratings of teacher effectivenss: Validity studies. American Educational Research Journal, 1971, 8, 435-445.

Meier, R. S., & Feldhusen, J. F. another look at Dr. Fox: Effect of stated purpose for evaluation, lecture expressiveness, and density of lecture content on student ratings. Journal of Educational Psychology, 1979, 71, 339-345.

Naftulin, D. H., Ware, J. E., & Donnelly, F. A. The Dr. Fox lecture: A paradigm of educational seduction. Journal of Medical Education, 1973, 48, 630-635.

Perry, R. P., Abrami, P. C., & Leventhal, L. Educational seduction: The effect of instructor expressiveness and lecture content on student ratings and achievement. Journal of Educational Psychology, 1979, 71, 107-116.

Perry, R. P., Abrami, P. C., Leventhal, L., & Check, J. Instructor reputation: An expectancy relationship involving student ratings and achievement. Journal of Educational Psychology, 1979, 71, 776-787.

Perry, R. P., & Leventhal, L. Learned helplessness in the college classroom: Are teacher behaviors involved. Proceedings, XXII International Congress of Psychology, 1980, Leipzig, East Germany.

Perry, R. P., Niemi, R. R., & Jones, K. Effect of prior teaching evaluations and lecture presentation on ratings of teaching performance. Journal of Educational Psychology, 1974, 66, 851-856.

Pohlman, J. T. A multivariate analysis of selected class characteristics and student ratings of instruction. Multivariate Behavioral Research, 1975, 10, 81-92.

Ramagali, H. J., & Greenwood, G. E. Another assessment of the Dr. Fox effect: Paired lecture comparisons of expressiveness and content. Journal of Educational Psychology, 1981, in press.

Remmers, H. H. The college professor as the student sees him. Purdue University Studies in Higher Education, 1929, 29, 75.

Rodin, M., & Rodin, B. Student evaluations of teachers. Science, 1972, 177, 1164-1166.

Sullivan, A. M., & Skanes, G. R. Validity of student evaluation of teaching and the characteristics of successful instructors. Journal of Educational Psychology, 1974, 66, 584-590.

Ware, J. E., & Williams, R. G. The Dr. Fox effect: A study of lecturer effectiveness and the validity of student ratings of instruction. Journal of Medical Education, 1975, 50, 149-156.

Ware, J. E., & Williams, R. G. Seeing through the Dr. Fox effect: A response to Frey. Instructional Evaluation, 1979, 3(2), 6-10.

Ware, J. E., & Williams, R. G. A reanalysis of the Dr. Fox experiments. Instructional Evaluation, 1980, 4(2), 15-19.

Werdell, P. R. Course and teacher evaluation. (2nd Ed.) Washington, D.C.: United States National Students Association, 1967.

Williams, R. G., & Ware, J. E. Validity of student ratings of instruction under different incentive conditions: A further study of the Dr. Fox effect. Journal of Educational Psychology, 1976, 68, 48-56.

Williams, R. G., & Ware, J. E. An extended visit with Dr. Fox: Validity of student ratings of instruction after repeated exposures to a lecturer. American Educational Research Journal, 1977, 14, 449-457.